Flash Memory Overview

Steven Swanson
Humanity processed 9 Zettabytes in 2008*
Welcome to the Data Age!

*http://hmi.ucsd.edu
Solid State Memories

- NAND flash
  - Ubiquitous, cheap
  - Sort of slow, idiosyncratic
- Phase change, Spin torque MRAMs, etc.
  - On the horizon
  - DRAM-like speed
  - DRAM or flash-like density
Bandwidth Relative to disk

PCIe-Flash (2012)  
PCIe-PCM (2010)  
PCIe-PCM (2013?)  
DDR Fast NVM (2016?)

5917x → 2.4x/yr  
7200x → 2.4x/yr
Disk Density

Areal Density Perspective
44 Years of Technology Progress

- 1st Thin Film Head
- 1st MR Head
- 1st GMR Head
- Travelstar 30GT Deskstar 40GV
- Ultrastar 73LZX 100% CGR
- Microdrive II

8.5 Million X Increase

IBM Disk Drive Products
Industry Lab Demos

IBM RAMAC (First Hard Disk Drive)

Production Year 60 70 80 90 2000 2010

Areal Density Megabits/in²

IBM

NVSL
Non-volatile Systems Laboratory

1 Tb/square inch
Hard drive Cost

- Today at newegg.com: $0.04 GB ($0.00004/MB)
- Desktop, 2 TB
Why Are Disks Slow?

• They have moving parts :-(
  – The disk itself and the head/arm

• The head can only read at one spot.

• High end disks spin at 15,000 RPM
  – Data is, on average, 1/2 an revolution away: 2ms
  – Power consumption limits spindle speed
  – Why not run it in a vacuum?

• The head has to position itself over the right “track”
  – Currently about 150,000 tracks per inch.
  – Positioning must be accurate with about 175nm
takes 3-13ms
Making Disks Faster

• Caching
  – Everyone tries to cache disk accesses!
  – The OS
  – The disk controller
  – The disk itself.

• Access scheduling
  – Reordering accesses can reduce both rotational and seek latencies
RAID!

- Redundant Array of Independent (Inexpensive) Disks
- If one disk is not fast enough, use many
  - Multiplicative increase in bandwidth
  - Multiplicative increase in Ops/Sec
  - Not much help for latency.
- If one disk is not reliable enough, use many.
  - Replicate data across the disks
  - If one of the disks dies, use the replica data to continue running and re-populate a new drive.
- Historical foot note: RAID was invented by one of the textbook authors (Patterson)
RAID Levels

• There are several ways of ganging together a bunch of disks to form a RAID array. They are called “levels”

• Regardless of the RAID level, the array appears to the system as a sequence of disk blocks.

• The levels differ in how the logical blocks are arranged physically and how the replication occurs.
RAID 0

• Double the bandwidth.
• For an n-disk array, the n-th block lives on the n-th disk.
• Worse for reliability
  – If one of your drives dies, all your data is corrupt-- you have lost every nth block.
RAID 1

- Mirror your data
- 1/2 the capacity
- But, you can tolerate a disk failure.
- Double the bandwidth for reads
- Same bandwidth for writes.
• Stripe your data across a bunch of disks
• Use one bit to hold parity information
  – The number of 1’s at corresponding locations across the drives is always even.
• If you lose on drive, you can reconstruct it from the others.
  Read and write all the disks in parallel.
The Flash Juggernaut
Flash is Fast!

Hard Drives
- Lat.: 7.1ms
- BW: 2.6MB/s
  - 1x
  - 1x

PCle-Flash 2007
- 68us
- 250MB/s
  - 104x
  - 96x

• Random 4KB Reads from user space
Flash Operations

Read

0V 1V

20V Erase

0V 0V

Program

5V

0V

20V

0V

5V

0V

20V

0V

0V

Floating Gate

Word Line

External Gate

n-Source

n-Source

Floating Gate

n-Drain

n-Drain

p-substrate

p-substrate
Organizing Flash Cells into Chips

Select Transistors

Data storage

One NAND chain

One block

One page
Organizing Flash Cells into Chips

- ~16K blocks/chip
- ~16-64Gbits/chip
Flash Operations

Page: 0 1 2 3 4 ...
Block 0
Block 1
Block 2 ...
Block n

Erase Blocks
Program Pages

SLC: Single Level Cell
- == 1 bit

MLC: Multi Level Cell
- == 2 bits

TLC: Triple Level Cell
- == 3 bits

SLC: Single Level Cell
MLC: Multi Level Cell
TLC: Triple Level Cell
Single-Level Cell

Endurance: 100,000 Cycles
Data retention: 10 years
Read Latency: 25us
Program Latency: 100-200us

Charge on the floating gate

== 1 bit
Multi-Level Cell (2 bits)

Endurance: 5000-10,000 Cycles
Data retention: 3-10 years
Read Latency: 25-37us
Program Latency: 600-1800us
Triple-level Cell (3bits)

Endurance: ~500-1000 Cycles
Data retention: 3 years
Read Time: 60-120us
Program Time: 500-6500us

= 3 bits
3D NAND

- SLC, MLC, and TLC NAND cells are $4F^2$ devices.
  - $1.33 - 4F^2$ per bit

- Higher densities require 3D designs
  - Samsung has demonstrated 24 layers
  - 2-4x density boost

- http://bcove.me/xz2o1af5
Flash Failure Mechanisms

• Program/Erase (PE) Wear
  – Permanent damaged to the gate oxide at each flash cell
  – Caused by high program/erase voltages
  – Damage causes charge to leak off the floating gate

• Program disturb
  – Data corruption caused by interference from programming adjacent cells.
  – No permanent damage
Making Disks out Flash Chips

Read Pages
Write Pages
Erase Blocks
Hierarchical addresses
PE Wear

Read
Write
Flat address space
No wear limitations
SSD Maintain a map between “virtual” logical block addresses and “physical” flash locations.
Writing more data...

When you overwrite data, it goes to a new location.
Flash Translation Layer (FTL)

User
• Logical Block Address

Flash
• Write pages in order
• Erase/Write granularity
• Wears out

FTL
• Logical → Physical map
• Wear leveling
• Power cycle recovery
Centralized FTL State

Map

<table>
<thead>
<tr>
<th>LBA</th>
<th>Physical Page Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Block 5 Page 7</td>
</tr>
<tr>
<td>2k</td>
<td>Block 27 Page 0</td>
</tr>
<tr>
<td>4k</td>
<td>Block 10 Page 2</td>
</tr>
</tbody>
</table>

Write Point

<table>
<thead>
<tr>
<th>101001011010001</th>
</tr>
</thead>
<tbody>
<tr>
<td>010100100101011</td>
</tr>
<tr>
<td>101010110101001</td>
</tr>
<tr>
<td>111111111111111</td>
</tr>
<tr>
<td>111111111111111</td>
</tr>
<tr>
<td>111111111111111</td>
</tr>
</tbody>
</table>

Block Info Table

<table>
<thead>
<tr>
<th>Block</th>
<th>Erased</th>
<th>Erase Count</th>
<th>Valid Page Count</th>
<th>Sequence Number</th>
<th>Bad Block Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>False</td>
<td>3</td>
<td>15</td>
<td>5</td>
<td>False</td>
</tr>
<tr>
<td>1</td>
<td>True</td>
<td>7</td>
<td>0</td>
<td>-</td>
<td>False</td>
</tr>
<tr>
<td>2</td>
<td>False</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>False</td>
</tr>
</tbody>
</table>

Next Sequence Number: 12
Read

1. Read Data at LBA 2k

2. Map

<table>
<thead>
<tr>
<th>LBA</th>
<th>Physical Page Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Block 5</td>
</tr>
<tr>
<td>2k</td>
<td>Block 27</td>
</tr>
<tr>
<td>4k</td>
<td>Block 10</td>
</tr>
</tbody>
</table>

3. Flash Operation